GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators

Year: 2020 | Citations: 203 | Authors: Dingfan Chen, Tribhuvanesh Orekondy, Mario Fritz

Abstract

The wide-spread availability of rich data has fueled the growth of machine learning applications in numerous domains. However, growth in domains with highly-sensitive data (e.g., medical) is largely hindered as the private nature of data prohibits it from being shared. To this end, we propose Gradient-sanitized Wasserstein Generative Adversarial Networks (GS-WGAN), which allows releasing a sanitized form of the sensitive data with rigorous privacy guarantees. In contrast to prior work, our approach is able to distort gradient information more precisely, and thereby enabling training deeper models which generate more informative samples. Moreover, our formulation naturally allows for training GANs in both centralized and federated (i.e., decentralized) data scenarios. Through extensive experiments, we find our approach consistently outperforms state-of-the-art approaches across multiple metrics (e.g., sample quality) and datasets.